Research Proposal

Master Thesis at the LMU Munich

Department of Statistics

Recognition of handwritten cards within an old Occitan dictionary project

Supervision:

- Prof. Dr. Christian Heumann
- Esteban Garces Arias (Institut für Statistik, LMU)
- Matthias Schöffel (Bayerische Akademie der Wissenschaften (BAdW) / LMU)

Keywords: OCR, Handwritten Text Recognition, Deep Learning, Multimodal models, NLP, Computer vision, Historical linguistics, old Occitan.

Problem Statement:

This thesis proposal is based on a dictionary project at the Bavarian Academy of Sciences (https://badw.de/en/the-academy.html) establishing the state of the art dictionary for old Occitan, one of the minor Romance languages. The *Dictionnaire de l'occitan médiéval* (DOM) covers the medieval period 1000–1550 (details, see https://dom.badw.de/das-projekt.html). Its modern counterpart is still spoken in southern France.

The initial work for this project started in 1960. At that time, the entire material was written on cards (581.000 in total), serving as a basis for the daily work since then. Within this material, the central information (source, context, inferred meaning, and additional information) of every word was recorded. As this language is a non-standardized language, unlike modern languages, graphical variants were collected as well because they serve as an essential source for analyzing the (de-)standardization processes over time. The relations between these variants and their respective lemmata were noted by hand in reference cards (see fig. 1), which now need to be put in digital format.

Many approaches in the field of Handwritten Text Recognition (HTR) have been developed in the last years, including deep learning techniques, transformer architectures, and multimodal methods. Besides commercial APIs such as Google Cloud Vision and AWS Textract, several (multi-lingual) open-source models such as Tesseract, SimpleHTR or OrigamiNet have provided good results while performing HTR [1,2,3,4].

Figure 1: Six examples of reference cards for old Occitan. Each card has the following structure: graphical variant → lemma (standardized dictionary entry)

Proposed tasks and research questions to be discussed in the thesis:

- Literature review of statistical methods and current deep learning methods for HTR, particularly CV/NLP methods.
- Comparative study of pre-trained models for HTR. Discuss advantages and disadvantages in terms of scalability, stability, fit, and complexity of the models. Evaluate the effect of limited data on performance.
- Extend the existing methods and propose an approach that improves their performance.
- Analyze the effect of enhancing inputs with image-related features over the handwritten recognition results.
- Identify relevant patterns and steps that are critical to the performance of the applied methods.
- Optional: Participation in *dhmuc-Gespräche* (monthly talks on Digital Humanities) at the BAdW to share ideas/achievements during the project.
- Programming can be in R or Python.

Qualifications:

- Comfortable with written and oral communication in English.
- Good understanding of machine learning and deep learning concepts (roughly the content of the lectures Predictive Modeling, Introduction to Deep Learning. Knowledge in the fields of NLP and Computer Vision is a plus).
- Good programming skills in R or Python and a deep learning framework (e.g., pytorch).

Contact:

In case you are interested, please contact us:

- Prof. Dr. Christian Heumann (chris@stat.uni-muenchen.de)
- Esteban Garces Arias (Esteban.GarcesArias@stat.uni-muenchen.de)

Literature:

[1] Chung, J. and Deltiel, T. (2019) A Computationally Efficient Pipeline Approach to Full Page Offline Handwritten Text Recognition. CVPR 2019.
[2] Grother, P. and Hanaoka, K. (2016) NIST Handwritten Forms and Characters Database (NIST Special Database 19). DOI: http://doi.org/10.18434/T4H01C
[3] Ingle, R. et al. (2019) A Scalable Handwritten Text Recognition System. ICDAR 2019.
[4] Yousef, M. and Bishop, T. (2020) OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by Learning to Unfold. CVPR 2020.